Data Clustering: Principal Components, Hopfield and Self-Aggregation Networks*

Chris H.Q. Ding
NERSC Division, Lawrence Berkeley National Laboratory
University of California, Berkeley, CA 94720. chqding@lbl.gov

Abstract

We present a coherent framework for data clustering. Starting with a Hopfield network, we show the solutions for several well-motivated clustering objective functions are principal components. For MinMaxCut objectives motivated for ensuring cluster balance, the solutions are the nonlinearly scaled principal components. Using scaled PCA, we generalize to multi-way clustering, constructing a self-aggregation network, where connection weights between different clusters are automatically suppressed while connection weights within same clusters are automatically enhanced.

1 Introduction

Principal component analysis (PCA) is widely adopted as an effective unsupervised dimension reduction method. PCA is extended in many different directions [Hastie and Stuetzle, 1989; Kramer, 1991; Lee and Seung, 1999; Scholkopf *et al.*, 1998; Collins *et al.*, 2001].

The main justification is that PCA uses singular value decomposition (SVD) which is the best low rank approximation in L_2 norm to original data due to Eckart-Young theorem. However, this results alone is inadequate to explain the effectiveness of PCA. Here, we provide a new derivation of PCA based on optimizing suitable clustering objective functions and show that principal components are actually cluster indicator vectors in clustering.

Hopfield network[Hopfield, 1982] provide a convenient framework of our study. In particular, the self-aggregation network proposed in this work uses Hebb rule to encode pattern vectors. One feature of Hopfield associative-memory networks is that it can be adopted to solve hard combinatorial problems[Haykin, 1998 2nd ed].

Another thread of this work is the spectral graph partitioning [Fiedler, 1973; Pothen *et al.*, 1990; Hagen and Kahng, 1992; Shi and Malik, 2000; Ding *et al.*, 2001a; 2001b; Ng *et al.*, 2001; Meila and Shi, 2001], which uses Laplacian matrix of a graph. This arises naturally for balancing the clusters (see §2.4). Our approach differs from others mainly in well-motivated clustering objective functions. A further point is the recognition that spectral graph clustering is embedded in the scaled PCA thus leading to self-aggregation networks.

This paper combines above three threads and develop a coherent framework for clustering. We begin with two-way clustering in §2 and generalize to multi-way clustering in §3.

2 Two-way clustering

We start by formulate two-way clustering as a Hopfield network and derive PCA as cluster indicator vectors.

2.1 Hopfield Network and PCA

Given n data points and properly defined similarity or association $w_{ij} = w_{ji} \geq 0$ between points i, j, we form a network (a weighted graph) G with w_{ij} as the connection between nodes i, j. We wish to partition it into two clusters $\mathrm{c}_1, \mathrm{c}_2$. The result of clustering can be represented by an indicator vector $\mathbf{q}$, where

$$q(i) = \begin{cases} 1 & \text{if} \quad i \in \mathrm{c}_1 \\ -1 & \text{if} \quad i \in \mathrm{c}_2 \end{cases} \tag{1}$$

Consider the clustering objective,

$$J_1 = \mathbf{q}^T W \mathbf{q} = s(\mathrm{c}_1, \mathrm{c}_1) + s(\mathrm{c}_2, \mathrm{c}_2) - 2s(\mathrm{c}_1, \mathrm{c}_2) \tag{2}$$

where $s(\mathrm{c}_1, \mathrm{c}_2) = \sum_{i \in \mathrm{c}_1, j \in \mathrm{c}_2} w_{ij}$ is the overlap between $\mathrm{c}_1, \mathrm{c}_2$; $s(\mathrm{c}_1, \mathrm{c}_1) = \sum_{i,j \in \mathrm{c}_1} w_{ij}$ is the within-cluster similarity of c_1, and analogously for $s(\mathrm{c}_2, \mathrm{c}_2)$.

Now we propose a *min-max clustering principle*: data points are grouped into clusters such that the overlap $s(\mathrm{c}_1, \mathrm{c}_2)$ between different clusters are minimized while within-cluster similarities $s(\mathrm{c}_1, \mathrm{c}_1)$, $s(\mathrm{c}_2, \mathrm{c}_2)$ are maximized

$$\min s(\mathrm{c}_1, \mathrm{c}_2), \quad \max s(\mathrm{c}_1, \mathrm{c}_1), \quad \max s(\mathrm{c}_2, \mathrm{c}_2) \tag{3}$$

These conditions can be simultaneously satisfied by maximizing the energy (objective function) J_1. Using Hopfield model [Hopfield, 1982], the solution is obtained by the update rule

$$q^{(t+1)}(i) = \mathrm{sgn}[\sum_j w_{ij} q^{(t)}(j)].$$

or in vector form $\mathbf{q}^{(t+1)} = \mathrm{sgn}[W\mathbf{q}^{(t)}]$. If one relaxes $q(i)$ from discrete indicators to continuous values in $(-1, 1)$, the solution $\mathbf{q}$ is given by

$$W\mathbf{q} = \lambda \mathbf{q}. \tag{4}$$

*Work supported by Department of Energy (Office of Science, through a LBNL LDRD) under contract DE-AC03-76SF00098.

Since the matrix entries in W are non-negative, the first principal eigenvector $\mathbf{q}_1$ has all positive (or all negative) entries. Thus the desired solution is $\mathbf{q}_2$.

2.2 Hopfield network for bipartite graph

Here we extend the Hopfield networks for clustering bipartite graph. An example of bipartite graph is a $m \times n$ term-document association matrix $B = (b_{ij})$, where each row represents a word, each column represents a document, and b_{ij} the counts of co-occurrence of row r_i and column c_j. We show that the solution for clustering indicators is precisely the Latent Semantic Indexing[Deerwester *et al.*, 1990].

We wish to partition the r-type nodes of R into two parts $\mathrm{R}_1, \mathrm{R}_2$ and simultaneously partition the c-type nodes of C into two parts $\mathrm{C}_1, \mathrm{C}_2$, based on the clustering principle of minimizing between-cluster association and maximizing within-cluster association. We use indicator vector $\mathbf{f}$ to determine how to split R into $\mathrm{R}_1, \mathrm{R}_2$: $f(i) = 1$, or -1 depending on $r_i \in \mathrm{R}_1$ or $r_i \in \mathrm{R}_2$. We use g to determine how to split C into $\mathrm{C}_1, \mathrm{C}_2$: $g(i) = 1$, or -1 depending on $c_i \in \mathrm{C}_1$ or $c_i \in \mathrm{C}_2$.

(For presentation purpose, we index the nodes such that nodes within same cluster are indexed contiguously. The clustering algorithms presented are independent to this assumption. Bold face lower case letters are vectors. Matrices are denoted by upper case letters.) Thus we may write $\mathbf{f} = \begin{pmatrix} \mathbf{f}^{(+)} \\ \mathbf{f}^{(-)} \end{pmatrix}$, $\mathbf{g} = \begin{pmatrix} \mathbf{g}^{(+)} \\ \mathbf{g}^{(-)} \end{pmatrix}$ With this indexing, the association matrix is

$$B = \begin{pmatrix} B_{\mathrm{R}_1,\mathrm{C}_1} & B_{\mathrm{R}_1,\mathrm{C}_2} \\ B_{\mathrm{R}_2,\mathrm{C}_1} & B_{\mathrm{R}_2,\mathrm{C}_2} \end{pmatrix} \tag{5}$$

It is convenient to convert the bipartite graph into an undirected graph. We follow standard procedure and combine the two types nodes to one by setting

$$\mathbf{q} = \begin{pmatrix} \mathbf{f} \\ \mathbf{g} \end{pmatrix}, \; W = \begin{pmatrix} 0 & B \\ B^T & 0 \end{pmatrix}, \tag{6}$$

This induces an undirected graph G, whose adjacency matrix is the symmetric weight matrix W.

Consider the following objective function,

$$\begin{aligned} J_2(\mathrm{C}_1, \mathrm{C}_2; \mathrm{R}_1, \mathrm{R}_2) &= \frac{1}{2}\mathbf{q}^T W \mathbf{q} \qquad (7) \\ &= s(\mathrm{R}_1, \mathrm{C}_1) + s(\mathrm{R}_2, \mathrm{C}_2) - s(\mathrm{R}_1, \mathrm{C}_2) - s(\mathrm{R}_2, \mathrm{C}_1) \end{aligned}$$

where $s(\mathrm{R}_1, \mathrm{C}_2) \equiv \sum_{r_i \in \mathrm{R}_1, c_j \in \mathrm{C}_2} b_{ij}$, the sum of association between R_1 and C_2, and $s(\mathrm{R}_1, \mathrm{C}_2)$ is the sum of association between R_1 and C_2. $s(\mathrm{R}_1, \mathrm{C}_2)$ and $s(\mathrm{R}_2, \mathrm{C}_1)$ should be minimized. $s(\mathrm{R}_1, \mathrm{C}_1)$ is the sum of association within cluster 1 (see Fig.1), and $s(\mathrm{R}_2, \mathrm{C}_2)$ is the sum of association within cluster 2. $s(\mathrm{R}_1, \mathrm{C}_1)$ and $s(\mathrm{R}_2, \mathrm{C}_2)$ should be maximized. These conditions are simultaneously satisfied by maximizing J_2.

We can write down the Hopfield network update rule for $\mathbf{q}$. If one relaxes $q(i)$ from discrete indicators to continuous values, the solution $\mathbf{q}$ satisfies Eq.(4). Now utilizing the explicit structures of W and $\mathbf{q}$, we have

$$\begin{pmatrix} 0 & B \\ B^T & 0 \end{pmatrix} \begin{pmatrix} \mathbf{f} \\ \mathbf{g} \end{pmatrix} = \lambda \begin{pmatrix} \mathbf{f} \\ \mathbf{g} \end{pmatrix}. \tag{8}$$

The solutions to this equation are the singular value decomposition (SVD) of B: $\{\mathbf{f}_i\}$ are left singular vectors and $\{\mathbf{g}_i\}$ are right singular vectors of the SVD of B,

$$B = \sum_{k=1}^{m} \mathbf{f}_k \lambda_k \mathbf{g}_k^T = F_m \Lambda_m G_m^T. \tag{9}$$

Again, since the matrix entries in W are non-negative, the first principal components $\mathbf{u}_1, \mathbf{v}_1$ have entries of same sign; thus the desired solutions are $\mathbf{u}_2, \mathbf{v}_2$. Note that this is also the SVD employed in LSI. We summarize these results in

Theorem 1. Principal components are solutions for clustering indicators for clustering undirect graphs and bipartite graphs under appropriate Hopfield network models.

2.3 Principal component clustering

In the objective J_1, we may explicitly enforce a balance of clusters $\mathrm{C}_1, \mathrm{C}_2$. From Eq.1, we need to minimize $|\sum_i q(i)|$. Thus we consider the clustering objective,

$$\tilde{J}_1 = \mathbf{q}^T W \mathbf{q} - \mu(\sum_i q(i))^2 = \mathbf{q}^T \left(W - \mu \mathbf{e}\mathbf{e}^T\right) \mathbf{q} \tag{10}$$

where $\mu > 0$ is a parameter and $\mathbf{e} = (1, \cdots, 1)^T$. We adjust μ to control the levele of balance between $\mathrm{C}_1, \mathrm{C}_2$. The principal eigenvector $\mathbf{q}_1$ of $\widetilde{W} = W - \mu \mathbf{e}\mathbf{e}^T$ is the desired indicator vector.

What is the reasonable choice of μ? A natural choice is to set $\mu = w_{..}/n^2, w_{..} = \sum_{ij} w_{ij}$, the average of w_{ij}. With this choice, the modified weight matrix satisfies the sum-to-zero condition:

$$\sum_{ij} \widetilde{w}_{ij} = \sum_{ij} (w_{ij} - w_{..}/n^2) = 0. \tag{11}$$

This condition can be further refined by centering each column and each rows, such that

$$\sum_i \overline{w}_{ij} = 0, \forall j; \quad \sum_j \overline{w}_{ij} = 0, \forall i. \tag{12}$$

where

$$\overline{w}_{ij} = w_{ij} - w_{i.}/n - w_{.j}/n + w_{..}/n^2 \tag{13}$$

(here column and row sums, $w_{i.} = \sum_j w_{ij}, w_{.j} = \sum_i w_{ij}$ and $w_{..}$ are standard notations in statistics.) Now the desired cluster indicator vector is the principal eigenvector of

$$\overline{W}\mathbf{q} = \lambda \mathbf{q}$$

The fully centered $\overline{W}$ has a useful property that all eigenvectors of $\overline{W}$ with nonzero eigenvalues have the sum-to-zero property: $\sum_i \bar{q}(i) = 0$. This is because (1) $\bar{\mathbf{q}}_0 = \mathbf{e}$ is an eigenvector of $\overline{W}$ with $\lambda_0 = 0$; (2) all other eigenvectors are orthogonal to $\bar{\mathbf{q}}_0$, i.e, $\bar{\mathbf{q}}^T \mathbf{e} = 0$.

The sum-to-zero condition $\sum_i \bar{q}(i) = 0$ does not necessarily imply that the sizes of the two cluster, n_1, n_2, should be equal. In fact, the cluster indicator vector should be refined to

$$\bar{q}(i) = \begin{cases} \sqrt{n_2/n_1} & \text{if } \; i \in \mathrm{C}_1 \\ -\sqrt{n_1/n_2} & \text{if } \; i \in \mathrm{C}_2 \end{cases} \tag{14}$$

(in this paper, all vectors are implicitly normalized to 1 using L_2 norm). Correspondingly, the clustering objective becomes

$$\bar{J}_1 = \bar{\mathbf{q}}^T \overline{W} \bar{\mathbf{q}} = n_1 n_2 \left[\frac{s(\mathrm{C}_1,\mathrm{C}_1)}{n_1^2} + \frac{s(\mathrm{C}_2,\mathrm{C}_2)}{n_2^2} - 2\frac{s(\mathrm{C}_1,\mathrm{C}_2)}{n_1 n_2} \right] \quad (15)$$

Clearly, the first two terms represent the average within-cluster similarities which are maximized, and the 3rd term represent average between-cluster similarities which are minimized. The factor $n_1 n_2$ encourages cluster balance, since $n_1 + n_2 = n$, and $\max(n_1 n_2)$ is reached when $n_1 = n_2 = n/2$. We summarize these results in

Theorem 2. Solution for clustering objective $\tilde{J}_1$ is given by $\tilde{\mathbf{q}}_1$; and solution for clustering objective $\bar{J}_1$ is $\bar{\mathbf{q}}_1$.

We call these clustering schemes with objective functions $J_1, \bar{J}_1, \tilde{J}_1$ as principal component clustering methods because of the use of principal components.

All above discussions apply to bipartite graphs. For example, $\bar{J}_1$ becomes

$$\bar{J}_2 = \sqrt{n_{\mathrm{R}_1} n_{\mathrm{R}_2} n_{\mathrm{C}_1} n_{\mathrm{C}_2}} \cdot \left[\frac{s(\mathrm{R}_1,\mathrm{C}_1)}{n_{\mathrm{R}_1} n_{\mathrm{C}_1}} + \frac{s(\mathrm{R}_2,\mathrm{C}_2)}{n_{\mathrm{R}_2} n_{\mathrm{C}_2}} - \frac{s(\mathrm{R}_1,\mathrm{C}_2)}{n_{\mathrm{R}_1} n_{\mathrm{C}_2}} - \frac{s(\mathrm{R}_2,\mathrm{C}_1)}{n_{\mathrm{R}_2} n_{\mathrm{C}_1}} \right] \quad (16)$$

where $n_{\mathrm{R}_1}, n_{\mathrm{R}_2}$ are sizes of row clusters ($n_{\mathrm{R}_1} + n_{\mathrm{R}_2}$ = total row size); and $n_{\mathrm{C}_1}, n_{\mathrm{C}_2}$ are sizes of column clusters ($n_{\mathrm{C}_1} + n_{\mathrm{C}_2}$ = total column size). Again, the first two terms represent the average within-cluster similarities which are maximized, and the last two terms represent average between-cluster similarities which are minimized.

We note that $\bar{J}_1$ is partially motived by the classic scaling (also called principal coordinate analysis) in statistics [Borg and Gronen, 1997]. Suppose we are given pairwise distances d_{ij} for a dataset, define the pairwise similarity as $w_{ij} = -\frac{1}{2}d_{ij}^2$, follow through the same centering procedures and solve for the eigenvectors of $\overline{W}$. The classic scaling theorem proves that if d_{ij} are the Euclidean distances, the first K eigenvectors of $\overline{W}$ will recover the original coordinates. This theorem justifies the use of the K principal eigenvectors as the coordinates in multidimensional scaling. However, our clustering approach does not emphasize the recovery of original coordinates, and the objective function $\bar{J}_1$ is well-motivated for clustering only.

2.4 Cluster balance and MinMaxCut

Principal component clustering of §2.3 is motivated for balancing clusters by adding a panelity term to J_1. Here we balance clusters through a divisive approach.

Maximizing J_1 of Eq.2 is equivalent to[1]

$$\min \frac{s(\mathrm{C}_1,\mathrm{C}_2)}{s(\mathrm{C}_1,\mathrm{C}_1) + s(\mathrm{C}_2,\mathrm{C}_2)} \quad (17)$$

Although $s(\mathrm{C}_1,\mathrm{C}_1) + s(\mathrm{C}_2,\mathrm{C}_2)$ are maximized, it often occurs that one cluster is much larger than another, $s(\mathrm{C}_1,\mathrm{C}_1) \gg s(\mathrm{C}_2,\mathrm{C}_2)$ or vice versa.

[1]Since $\exp(x)$ is monotonic, $\max J_1 \Rightarrow \max\exp(J_1)$, which is $\min \exp\left[s(\mathrm{C}_1,\mathrm{C}_2)\right] / \exp\left[s(\mathrm{C}_1,\mathrm{C}_1) + s(\mathrm{C}_2,\mathrm{C}_2)\right]$. This is approximately Eq.17.

To overcome this problem, we seek to prevent either $s(\mathrm{C}_1,\mathrm{C}_1)$ or $s(\mathrm{C}_2,\mathrm{C}_2)$ become very small. We optimize[Ding *et al.*, 2001b]

$$\min J_3, \quad J_3 = \frac{s(\mathrm{C}_1,\mathrm{C}_2)}{s(\mathrm{C}_1,\mathrm{C}_1)} + \frac{s(\mathrm{C}_1,\mathrm{C}_2)}{s(\mathrm{C}_2,\mathrm{C}_2)} \quad (18)$$

One can show that

$$\min_{\mathbf{q}} J_3(\mathbf{q}) \Rightarrow \min_{\mathbf{q}} \frac{\mathbf{q}^T(D-W)\mathbf{q}}{\mathbf{q}^T D \mathbf{q}}, \quad (19)$$

subject to $\mathbf{q}^T W \mathrm{e} = \mathbf{q}^T D \mathrm{e} = 0$, where $D = (w_{i.})$ is a diagonal matrix. We relax $q(i)$ from discrete indicators to real values in $(-1, 1)$. The solution of $\mathbf{q}$ for minimizing the Rayleigh quotient of Eq.(19) is given by

$$(D - W)\mathbf{q} = \lambda D \mathbf{q}, \quad (20)$$

which can be written as

$$W\mathbf{q} = \zeta D\mathbf{q},\ \zeta = 1 - \lambda. \quad (21)$$

Again, the desired solutions is the eigenvector associated with the second largest eigenvalue. Note that by comparing Eq.(21) with Eq.(4), the net effect for cluster balance is diagonal scaling. This diagonal scaling, however, leads to an important feature of self-aggregation (see §3.2).

2.5 Cluster balance for bipartite graphs

For balanced clustering of bipartite graphs, object function of Eq.(7) becomes

$$J_4 = \frac{s(\mathrm{R}_1,\mathrm{C}_2) + s(\mathrm{R}_2,\mathrm{C}_1)}{2s(\mathrm{R}_1,\mathrm{C}_1)} + \frac{s(\mathrm{R}_1,\mathrm{C}_2) + s(\mathrm{R}_2,\mathrm{C}_1)}{2s(\mathrm{R}_2,\mathrm{C}_2)} \quad (22)$$

Using the representation W and $\mathbf{q}$ of Eq.6, and similar derivation [Ding, 2003; Zha *et al.*, 2001], the solution for optimization of J_4 is also given by Eq.21. Let $D_r = \mathrm{diag}(Be)$ and $D_c = \mathrm{diag}(B^T e)$, we have

$$D = \begin{pmatrix} D_r & 0 \\ 0 & D_c \end{pmatrix}, \quad \mathbf{z} = \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix} = \begin{pmatrix} D_r^{1/2}\mathbf{f} \\ D_c^{1/2}\mathbf{g} \end{pmatrix}. \quad (23)$$

Substituting into Eq.(21), we have

$$\begin{pmatrix} 0 & \widehat{B} \\ \widehat{B}^T & 0 \end{pmatrix} \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix} = \zeta \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix}, \quad (24)$$

where

$$\widehat{B} = D_r^{-1/2} B D_c^{-1/2}. \quad (25)$$

The solutions to Eq.(24) are SVD of $\widehat{B}$: $\widehat{B} = \sum_k \mathbf{u}_k \lambda_k \mathbf{v}_k^T$; The clustering indicator vectors are $\mathbf{f}_k = D_r^{-1/2}\mathbf{u}_k$ for row objects and $\mathbf{g}_k = D_c^{-1/2}\mathbf{v}_k$ for column objects. Note that Eq.(24) is identical Eq.(8), with the corresponding relationship $B \Rightarrow \widehat{B}$, $\begin{pmatrix} \mathbf{f} \\ \mathbf{g} \end{pmatrix} \Rightarrow \begin{pmatrix} \mathbf{u} \\ \mathbf{v} \end{pmatrix}$. Therefore, the net effect of clustering balancing using MinMaxCut objective of Eq.(22) over the simple objective Eq.(7) is the scaling of the association matrix B in Eq.(25). However, with this scaling, the self-aggregation property emerges (see §3.4).

3 K-way clustering

The above mainly focus on 2-way clustering. Below we generalize to K-way clustering, $K \geq 2$. The generalization is based on the key observation that the solution for cluster indicator vectors (see §2.4 and §2.5) are scaled principal components, as we discuss next.

3.1 Scaled principal components

Associations among data objects are mostly quantified by a similarity metric. The scaled principal component approach starts with a nonlinear (non-uniform) scaling of W. Noting that $W = D^{1/2}(D^{-1/2}WD^{-1/2})D^{1/2}$, we apply spectral decomposition on the scaled matrix

$$\widehat{W} = D^{-1/2}WD^{-1/2}$$

instead of on W. This leads to

$$W = D^{1/2}(\sum_k \mathbf{z}_k \zeta_k \mathbf{z}_k^T)D^{1/2} = D\sum_k \mathbf{q}_k \zeta_k \mathbf{q}_k^T D \quad (26)$$

Here we call $\mathbf{q}_k = D^{-1/2}\mathbf{z}_k$ the *scaled principal components*. they are obtained by solving the eigenvalue system

$$\widehat{W}\mathbf{z} = (D^{-1/2}WD^{-1/2})\mathbf{z} = \zeta\mathbf{z}. \quad (27)$$

Note Eq.(27) is identical to Eq.(20), or Eq.(21), thus scaled principal components are cluster indicator vectors in MinMaxCut (see §2.3).

3.2 Self-aggregation network

In Hopfield networks, a pattern $\mathbf{q}_1$ is encoded into the network as $\mathbf{q}_1\mathbf{q}_1^T$ (the Hebb rule); multiple patterns are encoded additively. In our problem, a pattern is a cluster partitioning indicator vector. We define a self-aggregation network with the connection weights

$$W_{SA} = \mathbf{q}_1\mathbf{q}_1^T + \cdots + \mathbf{q}_k\mathbf{q}_k^T = Q_K Q_K^T. \quad (28)$$

$Q_K = (\mathbf{q}_1, \cdots, \mathbf{q}_K)$. Here we highlights several important properties of W_{SA} and provides several example applications. (Self-aggregation is first studied in [Ding *et al.*, 2001a].)

W_{SA} has an interesting self-aggregation property enforced by within-cluster association (connectivity). To prove, we apply perturbation analysis by writing $\widehat{W} = \widehat{W}^{(0)} + \widehat{W}^{(1)}$, where $\widehat{W}^{(0)}$ is the similarity matrix when clusters are well-separated (zero-overlap) and $\widehat{W}^{(1)}$ accounts for the overlap among clusters and is treated as a perturbation[Ding *et al.*, 2001a]. This perturbation approach is standard in quantum physics[Mathews and Walker, 1971].

Well-separated clusters

First, we consider the case where clusters are well separated, i.e., no overlap (no connectivity) exist between different clusters. Let $W = (W_{pq}), \quad p, q = 1, \cdots, K$. be the subblock decomposition of W. No overlap is represented by the fact that off-diagonal blocks are zero: $W_{pq} = 0, p \neq q$. Now, $W - D = \mathrm{diag}(W_{11} - D_{11}, \cdots, W_{KK} - D_{KK})$ The eigenvalue problem decouples into K independent system $(W_{kk} - D_{kk})\mathbf{q} = \lambda D_{kk}\mathbf{q}$. Clearly, $\mathbf{e}_k = (1, \cdots, 1)^T$ with the size of cluster G_k, is the eigenvector for G_k. Let $s_{pq} = \sum_{i\in G_p}\sum_{j\in G_q} w_{ij}$, and $C = (c_{k\ell})$ be an arbitrary orthonormal matrix. Then

$$\mathbf{q}_k = (c_{1k}\mathbf{e}_1/s_{11}^{1/2}, \cdots, c_{Kk}\mathbf{e}_K/s_{KK}^{1/2})^T, \quad k = 1, \cdots, K, \quad (29)$$

are the K eigenvectors with eigenvalue $\zeta_k = 1$. $\mathbf{q}_k$ is a linear combination of K step functions, i.e., piece-wise constant function. Clearly, all data objects within the same cluster have *identical* elements in $\mathbf{q}$. The coordinate of object i in the K-dim SPCA space is $\mathbf{r}_i = (q_1(i), \cdots, q_K(i))^T$. Thus objects within a cluster are located at (self-aggregate into) the same point in SPAC space.

Scaled principal components are not *unique*, since C could be any orthogonal matrix. However,

$$Q_K Q_K^T = \mathrm{diag}(\mathbf{q}_1\mathbf{q}_1^T/s_{11}, \cdots, \mathbf{q}_K\mathbf{q}_K^T/s_{KK}), \quad (30)$$

is unique. Thus, self-aggregation of cluster member is equivalent to the fact that $Q_K Q_K^T$ has a block diagonal structure, where elements within the same diagonal block all have the same value.

It happens that two objects i, j within the same cluster are not initially connected, ie., $w_{ij} = 0$. However, $(Q_K Q_K^T)_{ij}$ gets the same connection strength as any other pairs within the same cluster, i.e., SA network performs a transitive closure on each connected component. Thus, in SA network, the within-cluster connections are enhanced.

Clusters overlap

Consider the case when overlaps among different clusters exist. The overlaps are treated as a perturbation. **Theorem 3**. At the first order, the K scaled principal components and their eigenvalues have the form

$$\mathbf{q} = D^{-1/2}X_K\mathbf{y}, \ \lambda = 1 - \zeta,$$

where $\mathbf{y}$ and ζ satisfy the eigensystem $\Gamma\mathbf{y} = \lambda\mathbf{y}$. The matrix Γ has the form $\Gamma = \Omega^{-1/2}\ \bar{\Gamma}\ \Omega^{-1/2}$, where

$$\bar{\Gamma} = \begin{pmatrix} h_{11} & -s_{12} & \cdots & -s_{1K} \\ -s_{21} & h_{22} & \cdots & -s_{2K} \\ \vdots & \vdots & \cdots & \vdots \\ -s_{K1} & -s_{K2} & \cdots & h_{KK} \end{pmatrix} \quad (31)$$

$h_{kk} = \sum_{p,p\neq k} s_{kp}$ and $\Omega = \mathrm{diag}(s_{11}, \cdots, s_K)$. This analysis is accurate to order $||\widehat{W}^{(1)}||^2/||\widehat{W}^{(0)}||^2$ for eigenvalues and to order $||\widehat{W}^{(1)}||/||\widehat{W}^{(0)}||$ for eigenvectors. □

Several features of SPCA can be obtained from Theorem 3:

Corollary 3.1. SA network $Q_K Q_K^T$ has the same block diagonal form of Eq.(30), within the accuracy of Theorem 2. This assures that objects within the same cluster will self-aggregate as in Theorem 1.

Corollary 3.2. The first scaled principal component is $\mathbf{q}_1 = D^{-1/2}X_K\mathbf{y}_1 = (1, \cdots, 1)^T$ with $\lambda_1 = 0, \zeta_1 = 1$. λ_1 and $\mathbf{q}_1$ are also the exact solutions to the original Eq.(20).

Corollary 3.3. When $K = 2$, the second principal component is

$$\mathbf{q}_2 = D^{-1/2}X_2\mathbf{y}_2 = \sqrt{\frac{s_{22}}{s_{11}}}\,\mathbf{e}^{(1)} - \sqrt{\frac{s_{11}}{s_{22}}}\,\mathbf{e}^{(2)}. \quad (32)$$

Its eigenvalue is

$$\lambda_2 = \frac{s_{12}}{s_{11}} + \frac{s_{12}}{s_{22}}. \tag{33}$$

Note that this is precisely J_3 of Eq.18, the clustering objective we started with: our clustering framework is consistent.

Example 1. A dataset of 3 clusters with substantial random overlap between the clusters. All edge weights are 1. The similarity matrix and results are shown in Fig.1, where nonzero matrix elements are shown as dots. The exact λ_2 and approximate $\tilde{\lambda}_2$ from Theorem 3 are close:

$$\lambda_2 = 0.300, \ \tilde{\lambda}_2 = 0.268.$$

W_{SA} is much sharper than the original weight matrix W clearly due to self-aggregation: connections between different clusters are substantially suppressed while connections within same clusters are substantially enhanced.

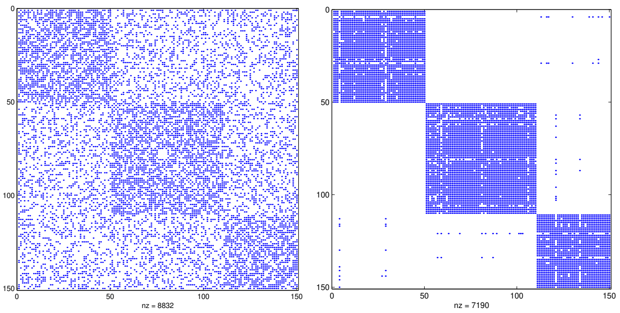

Figure 1: Left: similarity matrix W. Diagonal blocks represent weights inside clusters and off-diagonal blocks represent overlaps between clusters. Right: Computed W_{SA}.

Application 1. In DNA micro-array gene expression profiling, responses of thousands of genes from tumor tissues are simultaneously measured. We apply SPCA framework to gene expression profiles of lymphoma cancer samples[Alizadeh *et al.*, 2000]. Three cancer and three normal subtypes are shown in Fig.2. This is a difficult case due to the large variations of cluster sizes (the number of samples in each subtype are shown in parentheses in Fig.2B). Self-aggregation is evident in Figure 2B and 2C. The computed clusters correspond quite well to the normal and cancer subtypes identified by human experts.

3.3 Dynamic aggregation

The self-aggregation process can be repeated to obtain sharper clusters. W_K is the low-dimensional projection that contains the essential cluster structure. Combining this structure with the original similarity matrix, we obtain a new similarity matrix containing sharpened cluster information:

$$W^{(t+1)} = (1-\alpha)W_K^{(t)} + \alpha W^{(t)}, \tag{34}$$

where $W_K^{(t)} = DW_{\text{SA}}D$ for $W^{(t)}$, the weight matrix at t-th iteration, Setting $\lambda_k = 1$ is crucial for enforcing the cluster structure. $W^{(1)} = W$ and $\alpha = 0.5$.

Applying SA net on $W^{(2)}$ leads to further aggregation (see Figure 2C). The eigenvalues of the 1st and 2nd SA net are

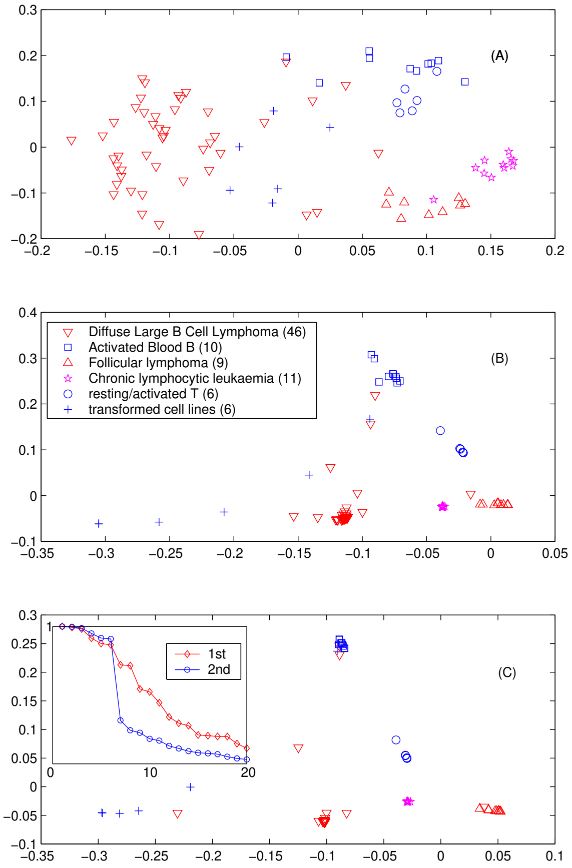

Figure 2: Gene expression profiles of cancerous and normal lymphoma tissues samples from Alizadeh et al. in original Euclidean space (A), in scaled PCA space (B), and in scaled PCA space after one iteration of Eq.34 (C). In all 3 panels, objects in original space are shown in 2D-view spanned by the first two PCA components. Cluster structures become clearer due to self-aggregation. The insert in (C) shows the eigenvalues of the 1st and 2nd SA network.

shown in the insert in Figure 1C. As iteration proceeds, a clear gap is developed, indicating that clusters becoming more separated.

Noise reduction

SA net has noises. For example, W_{SA} has sometimes negative weights $(W_{\text{SA}})_{ij}$ whereas we expect them to be nonnegative. However, by Corollaries 2.1 and 3.1, W_{SA} has a diagonal block structure and every elements in the block are identical (Eq.30) even when overlaps exit. This property allows us to interpret $W_{\text{SA}} = QQ^T$ as the probability that two objects i, j belong to the same cluster:

$$p_{ij} = (W_{\text{SA}})_{ij} / (W_{\text{SA}})_{ii}^{1/2} (W_{\text{SA}})_{jj}^{1/2}.$$

To reduce noise in the dynamic aggregation, we set

$$(W_K)_{ij} = 0 \quad \text{if} \quad p_{ij} < \beta, \tag{35}$$

where $0 < \beta < 1$ and we chose $\beta = 0.8$. Noise reduction is an integral part of SA net. In our experiments, final results are insensitive to α, β. The above dynamic aggregation repeats self-aggregation process and forces data objects move towards the attractors, which are the desired clusters and their principal eigenvalues approach 1 (see insert in Fig.2C). Usually, after one or two iterations the cluster structure becomes evident.

3.4 Bipartite graphs

SPCA applies to bipartite graph problems as well. The nonlinear scaling factors are D_r and D_c (cf. Eq.23). Let $B = D_r^{1/2}(\widehat{B})D_c^{1/2}$, where $\widehat{B}$ is given in Eq.(25). Applying SVD on $\widehat{B}$, we obtain

$$B = D_r^{1/2}(\sum_k \mathbf{u}_k \zeta_k \mathbf{v}_k^T) D_c^{1/2} = D_r \sum_k \mathbf{f}_k \zeta_k \mathbf{g}_k^T D_c. \quad (36)$$

Scaled principal components are $\mathbf{f}_k = D_r^{-1/2}\mathbf{u}_k$ for row objects and $\mathbf{g}_k = D_c^{-1/2}\mathbf{v}_k$ for column objects. They have the same self-aggregation and related properties. We note that SPCA on bipartite graph leads to correspondence analysis[Greenacre, 1984] from a new perspective.

The structure of self-aggregation network $Q_K Q_K^T$ can be meaningfully further decomposed:

$$Q_K Q_K^T = \begin{pmatrix} F_K F_K^T & F_K G_K^T \\ G_K F_K^T & G_K G_K^T \end{pmatrix}.$$

where $F_K = (\mathbf{f}_1, \cdots, \mathbf{f}_K)$ and $G_K = (\mathbf{g}_1, \cdots, \mathbf{g}_K)$. $F_K F_K^T$ provides the cluster structure for row objects, while $G_K G_K^T$ provides the cluster structure for column objects. The off-diagonal block matrix $F_K G_K^T$ provides the sharpened association between row and column objects[Ding, 2003].

4 Discussions

In this paper, we present a data clustering framework based on properly motivated Hopfield network, and show the clustering indicators are principal components. Further motivated by cluster balance, we extend the framework to MinMaxCut that utilized the Laplacian matrix of a graph. The framework is generalized to multi-way clustering, by using scaled principal components, and self-aggregation networks are constructed. We prove the cluster member self-aggregation property of the network. This framework extends naturally to bipartite graphs which leads to row-row, column-column and row-column SA nets that simultaneously cluster the row and column objects.

In self-aggregation, data objects move towards each other guided by connectivity. This is similar to the self-organizing map [Kohonen, 1989], where feature vectors self-organize into a 2D feature map while data objects remain fixed. All these have a connection to recurrent networks [Hopfield, 1982; Haykin, 1998 2nd ed]. In Hopfield network, features are stored as associative memories. In more complicated networks, connection weights are dynamically adjusted to learn or discover the patterns. The self-aggregation network provides a new mechanism to realize this unsupervised learning.

References

[Alizadeh *et al.*, 2000] A.A. Alizadeh, M.B. Eisen, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403:503–511, 2000.

[Borg and Gronen, 1997] I. Borg and P. Gronen. *Modern multidimensional scaling: theory and applications*. Springer Verlag, 1997.

[Collins *et al.*, 2001] M. Collins, S. Dasgupta, and R. Schapire. A generalization of principal component analysis to the exponential family. *Neural Info. Processing Systems (NIPS 2001)*, 2001.

[Deerwester *et al.*, 1990] S. Deerwester, S.T. Dumais, T.K. Landauer, G.W. Furnas, and R.A. Harshman. Indexing by latent semantic analysis. *J. Amer. Soc. Info. Sci*, 41:391–407, 1990.

[Ding *et al.*, 2001a] C. Ding, X. He, and H. Zha. A spectral method to separate disconnected and nearly-disconnected web graph components. In *Proc. ACM Int'l Conf Knowledge Disc. Data Mining (KDD)*, pages 275–280, 2001.

[Ding *et al.*, 2001b] C. Ding, X. He, H. Zha, M. Gu, and H. Simon. A min-max cut algorithm for graph partitioning and data clustering. *Proc. IEEE Int'l Conf. Data Mining*, pages 107–114, 2001.

[Ding *et al.*, 2002] C. Ding, X. He, H. Zha, and H. Simon. Unsupervised learning: self-aggregation in scaled principal component space. *Proc. 6th European Conf. Principles of Data Mining and Knowledge Discovery (PDKK 2002)*, pages 112–124, 2002.

[Ding, 2003] C. Ding. Document retrieval and clustering: from principal component analysis to self-aggregation networks. *Int'l Workshop on AI and Statistics*, pages 78–85, 2003.

[Fiedler, 1973] M. Fiedler. Algebraic connectivity of graphs. *Czech. Math. J.*, 23:298–305, 1973.

[Greenacre, 1984] M. J. Greenacre. *Theory and Applications of Correspondence Analysis*. Academic press, 1984.

[Hagen and Kahng, 1992] L. Hagen and A.B. Kahng. New spectral methods for ratio cut partitioning and clustering. *IEEE. Trans. on Computed Aided Desgin*, 11:1074–1085, 1992.

[Hastie and Stuetzle, 1989] T. Hastie and W. Stuetzle. Principal curves. *J. Amer. Stat. Assoc*, 84:502–516, 1989.

[Haykin, 1998 2nd ed] S.S. Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice Hall, 1998, 2nd ed.

[Hopfield, 1982] J.J. Hopfield. Neural networks and physical systems with emergent collective computation abilities. *Proc. Nat'l Acad Sci USA*, 79:2554–2558, 1982.

[Kohonen, 1989] T. Kohonen. *Self-organization and Associative Memory*. Springer-Verlag, 1989.

[Kramer, 1991] M.A. Kramer. Nonlinear principal component analysis using autoassociative neural networks. *AIChE Journal*, 37:233–243, 1991.

[Lee and Seung, 1999] D. D. Lee and H. S. Seung. Learning the parts of objects with nonnegative matrix factorization. *Nature*, 401:788–791, 1999.

[Mathews and Walker, 1971] J. Mathews and R.L. Walker. *Mathematical Methods of Physics*. Addison-Wesley, 1971.

[Meila and Shi, 2001] M. Meila and J. Shi. A random walks view of spectral segmentation. *AI-statistics Workshop*, 2001.

[Ng *et al.*, 2001] A.Y. Ng, M.I. Jordan, and Y. Weiss. On spectral clustering: Analysis and an algorithm. *Proc. Neural Info. Processing Systems (NIPS 2001)*, 2001.

[Pothen *et al.*, 1990] A. Pothen, H. D. Simon, and K. P. Liou. Partitioning sparse matrices with egenvectors of graph. *SIAM Journal of Matrix Anal. Appl.*, 11:430–452, 1990.

[Scholkopf *et al.*, 1998] B. Scholkopf, A. Smola, and K. Muller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998.

[Shi and Malik, 2000] J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE. Trans. on Pattern Analysis and Machine Intelligence*, 22:888–905, 2000.

[Zha *et al.*, 2001] H. Zha, X. He, C. Ding, M. Gu, and H.D. Simon. Bipartite graph partitioning and data clustering. *Proc. 10th Int'l Conf. Information and Knowledge Management (CIKM 2001)*, pages 25–31, 2001.